

Homework 2 Solutions

Note: These solutions are not necessarily model answers. Rather, they are designed to be tutorial in nature, and sometimes contain a little more explanation than an ideal solution. Also, bear in mind that there may be more than one correct solution. The maximum total number of points available is 35.

1. For $i = 1, 2, \dots, 2^{20} - 2$, let X_i be an indicator random variable that assumes the value 1 if the 3 letters beginning at the i 'th letter in the text form the sequence "ape", and 0 otherwise. Clearly, $\Pr[X_i = 1] = 26^{-3}$, so $E[X_i] = 26^{-3}$. Now, let the random variable X be the number of times the sequence "ape" appears in the text. Then, 5pts

$$X = X_1 + X_2 + \dots + X_{2^{20}-2}.$$

By linearity of expectation, the expected number of times the sequence "ape" appears in the text is

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_{2^{20}-2}] = \frac{2^{20} - 2}{26^3} \approx 59.7.$$

2. (a) The distribution of X is given by 2pts

$$\begin{aligned} \Pr[X = t] &= \Pr[\text{exactly } k - 1 \text{ of the first } t - 1 \text{ tosses land heads}] \cdot \Pr[\text{the } t\text{'th toss lands heads}] \\ &= \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k} \cdot p \\ &= \binom{t-1}{k-1} p^k (1-p)^{t-k}. \end{aligned}$$

- (b) Let Y_1 be the number of tosses up to, and including, the first head. For $i = 2, \dots, k$, let Y_i be the number of tosses up to, and including, the i 'th head, starting immediately after the $i - 1$ 'th head. Then clearly $X = Y_1 + Y_2 + \dots + Y_k$. Each of Y_1, \dots, Y_k is a geometric r.v. with mean p , so from class $E[Y_1] = E[Y_2] = \dots = E[Y_k] = \frac{1}{p}$. Hence $E[X] = \frac{k}{p}$. 3pts

3. (a) Andrew's scheme involves repeatedly flipping 1000 coins until exactly 500 heads and 500 tails are obtained. The probability of this happening is $p = \binom{1000}{500} \frac{1}{2^{1000}} \sim \frac{1}{\sqrt{500\pi}}$, using Stirling's approximation as explained below. The number of trials until we succeed is a geometric r.v. with parameter p , so the expected number of trials is $\frac{1}{p} \sim \sqrt{500\pi}$, giving a total of $1000\sqrt{500\pi} \approx 39600$ coin flips. 3pts

It remains to explain the calculation with Stirling's formula. Note that, for any n , we have

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \sim \frac{(2n/e)^{2n} \sqrt{4\pi n}}{(n/e)^{2n} 2\pi n} = \frac{2^{2n}}{\sqrt{\pi n}}.$$

Plugging in $n = 500$ we get $\binom{1000}{500} \approx \frac{2^{1000}}{\sqrt{500\pi}}$, which is exactly what we used in the previous paragraph.

- (b) Betty's scheme involves tossing coins until either 500 heads or 500 tails have been obtained, and then padding the sequence to get 1000 tosses with exactly 500 heads and 500 tails. This is a bad scheme because it does not give a uniform distribution. To see this, consider for example the sequence consisting of 500 heads followed by 500 tails. Since there are $\binom{1000}{500}$ valid sequences, this particular one should have probability $1/\binom{1000}{500} \approx \sqrt{500\pi} 2^{-1000}$ using Stirling's approximation as in part (a). But in Betty's scheme, this sequence has probability 2^{-500} (i.e., the chance of getting 500 heads in a row), which is clearly way too high since $\sqrt{500\pi} \ll 2^{500}$. 3pts

- (c) A better scheme involves randomly selecting the positions of the 500 heads in the sequence of 1000 tosses. One way to do this is to repeatedly choose random positions in the range 1 to 1000 until 500 *distinct* random positions have been obtained. A random number in the range 1 to 1024 can be picked using 10 coin tosses (since $2^{10} = 1024$). So say we're picking our k^{th} position, where $1 \leq k \leq 500$. The probability that a random number in the range 1 to 1024 has not been previously selected and is ≤ 1000 is $\frac{1001-k}{1024} \geq \frac{501}{1024} \geq \frac{1}{2.05}$, so the expected number of trials (each involving 10 coin tosses) till such a value is found is ≤ 2.05 . The expected total number of coin tosses is then at most $\sum_{k=1}^{500} 2.05 \times 10 = 10250$. (Actually, we have been quite pessimistic here in taking the worst case over k . The exact expectation is $10 \times \sum_{k=1}^{500} \frac{1024}{1001-k} = 10240(\frac{1}{501} + \frac{1}{502} + \dots + \frac{1}{1000}) \approx 10240 \ln(\frac{1000}{501}) \approx 7100$.)

An even better method is the following (but we did not necessarily expect anybody to come up with this!). Suppose our first random coin toss is H. Then for our next toss, we should use a biased coin with heads probability 499/999. (Why? Think about the conditional probabilities.) By the same argument, if the first i flips have h_i heads (and $i - h_i$ tails), then at the next step we should flip a coin with the probability of heads being $(500 - h_i)/(1000 - i)$, and so on. But how do we simulate these biased coins with our fair coin? We claim that *any* bias p can be realized using an expected number of only two fair coin flips!!! To see this, write $p = 0.p_1p_2\dots$ in binary. Now generate a random number $r = r_1r_2\dots$ between 0 and 1 by successively choosing each binary digit r_i using an independent fair coin flip. We can stop when we know that $r > p$ or that $r < p$ (corresponding to an outcome of H or T respectively for our biased coin). And when do we know this? It is when we reach the first i for which $r_i \neq p_i$. But the expected number of tosses for this event to happen is just 2. (Why?) Putting all this together, the expected number of coin flips needed to generate the entire random sequence of 1000 flips is only 2000.

4. (a) Fix any $r \in \{1, 2, \dots, n\}$. The algorithm outputs r if and only if c_n, \dots, c_{r+1} land tails, and c_r lands heads. This happens with probability

$$\frac{1}{r} \cdot \prod_{j=n}^{r+1} \left(1 - \frac{1}{j}\right) = \frac{1}{r} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdot \dots \cdot \frac{r}{r+1} = \frac{1}{n}.$$

Hence, the scheme generates $r \in \{1, 2, \dots, n\}$ u.a.r.

- (b) Following the hint, suppose we pick each s_i using the method of part (a). (This thought experiment is OK because we know from part (a) that this is a valid way of generating an integer u.a.r.) Following through the whole process of generating the s_i , we see that it is equivalent to flipping each coin c_j , $j = n, n-1, \dots, 1$ until it lands tails, and then moving on to the next coin. The value m_j is the number of times c_j lands heads before it lands tails. The probability that c_j lands heads exactly m_j times before landing tails is $\left(\frac{1}{j}\right)^{m_j} \left(1 - \frac{1}{j}\right)$. Hence, the probability we generate the sequence (s_i) is

$$\prod_{j=2}^n \left(\frac{1}{j}\right)^{m_j} \left(1 - \frac{1}{j}\right).$$

(Note that $j = 1$ is missed out because $m_j = 1$ always.)

- (c) Write $r \in \{1, \dots, n\}$ in factored form as $\prod_p p^{\beta_p}$. Observe that

$$\Pr\left[m_p = \beta_p \text{ for all primes } p \leq n\right] = \prod_p \left(\frac{1}{p}\right)^{\beta_p} \left(1 - \frac{1}{p}\right) = \frac{\alpha_n}{r},$$

and thus the algorithm outputs r with probability $\frac{r}{n} \cdot \frac{\alpha_n}{r} = \frac{\alpha_n}{n}$.

(d) The probability that the algorithm does not fail is $\sum_{r=1}^n \Pr[\text{algorithm outputs } r] = n \cdot \frac{\alpha_n}{n} = \alpha_n$. 2pts
Hence, the expected number of trials is $\alpha_n^{-1} \sim 1.8 \ln n$.

(e) Let X denote the number of primality tests performed by one trial of the algorithm. We can write this as $X = X_1 + \dots + X_n$, where X_j is the indicator r.v. of the event that j is tested for primality. Note that this happens iff $m_j \geq 1$, so by our analysis in part (b) we have $E[X_j] = \Pr[j \text{ is tested}] = \frac{1}{j}$. Hence $E[X] = \sum_{j=1}^n E[X_j] = H_n$, as required. 2pts

(f) Let the r.v. Z denote the total number of primality tests performed. Then $Z = Z_1 + Z_2 + \dots + Z_T$, 3pts
where Z_t denotes the number of tests performed in the t th trial and T is the number of trials. (Note that the number of terms in this summation is T , which is a random variable!) Clearly $E[Z_t] = H_n$ for all t from part (e), and from part (d) we have $E[T] = \alpha_n^{-1}$. Now it is tempting to say that, since Z is the sum of T r.v.'s each with the same expectation $E[Z_1]$, then $E[Z] = E[T]E[Z_1] = H_n/\alpha_n$; this turns out to be the correct answer (see below), but we need to do some work to justify it. If the Z_t were independent of T , we would have a very simple argument (see Q1 of HW3). In this case, we don't have independence and we need to do a bit more work.¹

So, let us yet again use the trick of indicator r.v.'s and write $Z = \sum_{t=1}^{\infty} \sum_{j=1}^n Y_{t,j}$, where $Y_{t,j}$ takes on the value 1 if j is tested for primality on the t th trial, and 0 otherwise. Writing A_t for the event that we perform t or more trials, we have

$$\begin{aligned} E[Y_{t,j}] &= \Pr[j \text{ is tested in trial } t] \\ &= \Pr[A_t] \cdot \Pr[j \text{ is tested in trial } t \mid A_t] \\ &= (1 - \alpha_n)^{t-1} \cdot \frac{1}{j}. \end{aligned}$$

Therefore, by linearity of expectation,

$$\begin{aligned} E[Z] &= \sum_{t=1}^{\infty} \sum_{j=1}^n E[Y_{t,j}] \\ &= \sum_{t=1}^{\infty} \sum_{j=1}^n (1 - \alpha_n)^{t-1} \cdot \frac{1}{j} \\ &= \sum_{t=1}^{\infty} (1 - \alpha_n)^{t-1} \cdot \sum_{j=1}^n \frac{1}{j} \\ &= \frac{H_n}{\alpha_n}. \end{aligned}$$

Since H_n and α_n^{-1} are both $\Theta(\log n)$, we get that $E[Z] = \Theta(\log^2 n)$, as required.

¹A much more general argument can be obtained using the machinery of *martingales* (see MU, Chap. 12). We will discuss martingales later in the class if there is time.